August 11, 2020

Dear Client –

After reviewing your data file, I’ve drafted a table that outlines issues with that data that should be addressed prior to taking additional action on your data. The header of the table contains the table names, whereas the first column contains the Standard Data Quality dimensions used to evaluate a dataset.

I think a short meeting to discuss the data issues would be helpful, but let me go over each dimension which should allow you to quickly understand my concerns with that particular column.

The Accuracy dimension will list a field that has correct values for that field. Fir the product\_ifrst\_sold\_date of the Transactions dataset, the values listed are not in date format.

The Completeness dimension lists fields that contain missing values. There needs to be a decision whether entries with a missing value for any of the columns should be removed or we can ignore the null/missing value for calculations.

The Currency dimensions lists fields with values that are current. I did not list any fields for that dimension, but one example of where it might be relevant would be the property\_valuation field in the CustomerAddress dataset. The data seems to come from 2017 and the current year is 2020. Performing data analysis on past data has its value, but understanding the context in which the analysis is being done will help determine if that field is no longer current and new property\_valuation values would need to be researched.

The Relevancy dimension contains fields that might be irrelevant to the analysis, therefore should be considered in the analysis, data gathering and cleanup. I will refer back to the property\_valuation field in the CustomerAddress data set and ask the question, why would you be interested in the property valuation of your customers? The same could be said of the job\_title and job\_industry\_category fields in the New Customer List data set that you provided. There needs to be clarity in what value we expect to gain from those fields.

The Validity dimension contains fields with values that contain values that do not apply. An example of this is the gender field in the Customer Demographic dataset. That field contains the values - 'F', 'Male', 'Female', 'U', 'Femal', 'M'. We can understand ‘F’ and ‘Female’ refer to the same gender, but the value ‘U’ would be an example of an invalid value.

The Duplicates dimension identifies datasets which contain rows of data with duplicate values. No duplicates have been identified in any of the datasets provided to me.

Regards,

Steve Correa

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Transactions** | **Customer Address** | **Customer Demographic** | **New Customer List** |
| **Accuracy (Correct values)** | - product\_first\_sold\_date | - Address | - default |  |
| **Completeness (Data fields with values)** | - online\_order - brand - product\_line - product\_class - product\_size - standard\_cost - product\_first\_sold\_date |  | - last\_name - DOB - job\_title - job\_industry\_category - default - tenure | - last\_name - DOB - job\_title - job\_industry\_category |
| **Consistency (Values free from contradiction)** | - list\_price (no dollar symbol) - standard\_cost (dollar symbol) - cust\_id has values that does not exist in dataset Customer, column cust\_id | - State  (some values are abbreviated) | - gender - deceased\_indicator  (uses N/Y) - owns\_car  (uses No / Yes) |  |
| **Currency (Values up to date)** |  |  |  |  |
| **Relevancy (Data items with value / meta-data)** | - standard\_cost  (purpose of this monetary value) - list\_price  (purpose of this monetary value) | - property\_valuation | - default - tenure | - job\_title - job\_industry\_category - wealth\_segment - deceased\_indicator - tenure - property\_valuation - Rank (with respect to what) - Value (Is this home value?) |
| **Validity (Data containing allowable values)** | - product\_first\_sold\_date  (float value, no date values) |  | - gender |  |
| **Duplicates (Records that are duplicated)** |  |  |  |  |